New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG IOP Institute of Physics

Trojan-horse attacks threaten the security of practical quantum cryptography

To cite this article: Nitin Jain *et al* 2014 *New J. Phys.* **16** 123030

New Journal of Physics
The open access journal at the forefront of physics
Deutsche Physikalische Gesellschaft DPG | IOP Institute of Physics

Trojan-horse attacks threaten the security of practical quantum cryptography

Nitin Jain[1,3], Elena Anisimova[2], Imran Khan[1,3], Vadim Makarov[2], Christoph Marquardt[1,3] and Gerd Leuchs[1,3]

[1] Max Planck Institute for the Science of Light, Günther-Scharowsky-Str. 1/Bau 24, D-91058 Erlangen, Germany
[2] Institute for Quantum Computing, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada
[3] Friedrich-Alexander University Erlangen-Nürnberg (FAU), Institute for Optics, Information and Photonics, Staudtstrasse 7/B2, D-91058 Erlangen, Germany
E-mail: nitin.jain@mpl.mpg.de

Received 22 June 2014, revised 23 September 2014
Accepted for publication 14 October 2014
Published 10 December 2014

New Journal of Physics **16** (2014) 123030

doi:10.1088/1367-2630/16/12/123030

Abstract

A quantum key distribution (QKD) system may be probed by an eavesdropper Eve by sending in bright light from the quantum channel and analyzing the back-reflections. We propose and experimentally demonstrate a setup for mounting such a Trojan-horse attack. We show it in operation against the quantum cryptosystem Clavis2 from ID Quantique, as a proof-of-principle. With just a few back-reflected photons, Eve discerns Bob's (secret) basis choice, and thus the raw key bit in the Scarani–Acín–Ribordy–Gisin 2004 protocol, with higher than 90% probability. This would clearly breach the security of the cryptosystem. Unfortunately, Eve's bright pulses have a side effect of causing a high level of after-pulsing in Bob's single-photon detectors, resulting in a large quantum bit error rate that effectively protects this system from our attack. However, in a Clavis2-like system equipped with detectors with less-noisy but realistic characteristics, an attack strategy with positive leakage of the key would exist. We confirm this by a numerical simulation. Both the eavesdropping setup and strategy can be generalized to attack most of the current QKD systems, especially if they lack proper safeguards. We also propose countermeasures to prevent such attacks.

Keywords: quantum hacking, quantum key distribution, quantum cryptography, Trojan horse, security proofs, reflectometry

1. Introduction

Quantum key distribution (QKD) provides a method to solve the task of securely distributing symmetric keys between two parties Alice and Bob [1–3]. The security of QKD is based on the principles of quantum mechanics: an adversary Eve attempting to eavesdrop on the quantum key exchange inevitably introduces errors that warn Alice and Bob about her presence. In the last decade however, several vulnerabilities and loopholes in the physical implementations of QKD have been discovered, and proof-of-principle attacks exploiting them have shown the possibilities that Eve may get hold of the secret key (in part or full) without alerting Alice and Bob [4–9].

In most cases, vulnerabilities and loopholes arise due to technical imperfections or deficiencies of the hardware. For instance, no optical component can *perfectly* transmit, or *completely* absorb light. An optical pulse launched into a network of optic and optoelectronic components, e.g., a QKD system, encounters several sites of Fresnel reflection and Rayleigh scattering. Some light thereby travels opposite to the propagation direction of the input optical signal. The properties and functionality of some component inside a QKD system may thus be probed from the quantum channel by sending in sufficiently-bright light and analyzing the back-reflected light. This forms the basis of a Trojan-horse attack [10].

Neither the concept, nor the danger of a Trojan-horse attack on QKD systems is new [11, 12]. Also, it is the Alice subsystem that is typically considered vulnerable to this kind of attack since it prepares the quantum state in most QKD schemes. If a QKD system is operating, e.g., the Bennett–Brassard 1984 (BB84) protocol [1], then by sending a suitably-prepared bright pulse inside Alice and analyzing its back-reflections, Eve could obtain information about the setting of the device, such as the polarizer or the phase modulator (PM), responsible for encoding the secret bit.

A simple way for Alice to detect a Trojan-horse attack red-handed is to install a passive monitoring device. This is usually implemented by a suitable detector (or an array of detectors) that measures the incoming signal and raises an alarm whenever certain pre-characterized thresholds are crossed. However, this countermeasure cannot be straightforwardly adopted for the Bob subsystem since a passive monitoring device would introduce unwanted attenuation in the *already-quite-weak* states of light coming from the quantum channel and bring the secret key rates down further. Even more, it may not be able to provide the security as expected [13, 14]. Another countermeasure is to add an optical isolator to block the bright Trojan pulse from entering [10, 12, 15, 16]; however, this is not applicable to two-way systems such as plug-and-play schemes [17]. Even otherwise, the limitations of isolators if Eve resorts to an attack at wavelengths in the vicinity of 1300 nm or 1700 nm have recently been highlighted [13].

For the BB84 protocol, this does not pose a problem as Bob publicly declares his basis choice, i.e., the setting of his polarizer/PM. However, in the Scarani–Acín–Ribordy–Gisin 2004 (SARG04) protocol [18, 19], the secret bit is given by Bob's basis choice. If Eve can surreptitiously read Bob's PM setting (= 0 or $\pi/2$) from the quantum channel via a Trojan-horse attack, then she acquires knowledge of the raw key [20]. She can then apply the same

operations (sifting, error correction and privacy amplification [2, 3]) as Alice and Bob—in other words, eavesdrop on the secret key without being discovered.

SARG04 is more robust than BB84 against photon-number-splitting attacks [9, 21], which is useful for QKD systems such as Clavis2 [22] that employ attenuated laser sources. In the following sections however, we show that it can be vulnerable to Trojan-horse attacks on Bob. We believe this is the first proof-of-principle demonstration of such an attack on a practical QKD system (although static phase readout in Alice has been demonstrated before [10, 12], the previous experiments were not real-time and did not analyze the complete system). Furthermore, both our eavesdropping setup and strategy are universal: with simple modifications, they could be applied against entanglement-based, continuous-variable, or even the very recent measurement-device-independent QKD systems [23–26] if they lack proper safeguards against Trojan-horse attacks [13, 27]. In such cases, it may even be used to break the BB84 protocol.

2. Theory and preparatory measurements

To prepare for a practical Trojan-horse attack, the eavesdropper Eve needs to know the answers to (at least) the following questions:

(i) What time should a Trojan-horse pulse be launched by Eve into Bob?
(ii) What time would a back-reflected pulse of interest exit Bob and arrive on the quantum channel? And with what amplitude?
(iii) What properties may be analyzed in a back-reflected pulse?
(iv) How to avoid being detected by Alice and Bob?
(v) What is the most suitable wavelength for attack?

These questions are closely interrelated, and the answers to them naturally depend on the QKD system under attack. In sections 2.1–2.5 below, we address them specifically for Clavis2, the plug-and-play QKD system from ID Quantique; or to be more precise, with the aim of crafting and executing an attack on Clavis2-Bob while it runs SARG04. Figure 1(a) shows the basic scheme of the attack while figure 1(b) shows the optical schematic of Clavis2 that operates in a two-way configuration based on the plug-and-play principle [17]. We briefly describe the principle below, and in the appendix we discuss several (technical) details via a numerical simulation.

Bob contains both the laser and the detectors; he sends bright pulse pairs to Alice who prepares the quantum states and sends them back to Bob. For this, she randomly modulates the relative phase $\varphi_A = \{0, \pi/2, \pi, 3\pi/2\}$ between the optical modes of each pair, and applies an attenuation so that the mean photon number of the resultant weak coherent pulses (returning to Bob on the quantum channel) is as dictated by the protocol. For SARG04, the optimal value is $\mu_{\mathrm{SARG04}} = 2\sqrt{T}$, where T is the channel transmission [19]. Bob applies a binary modulation chosen randomly per pair ($\varphi_B = 0$ or $\pi/2$, corresponding to the secret bit 0_B or 1_B respectively) and his pre-calibrated [8] gated detectors measure Alice's quantum states. The actual transmission uses the concept of *frames*, a train of pulses that entirely fit in Alice's delay line in order to prevent errors that would otherwise result from Rayleigh backscattering

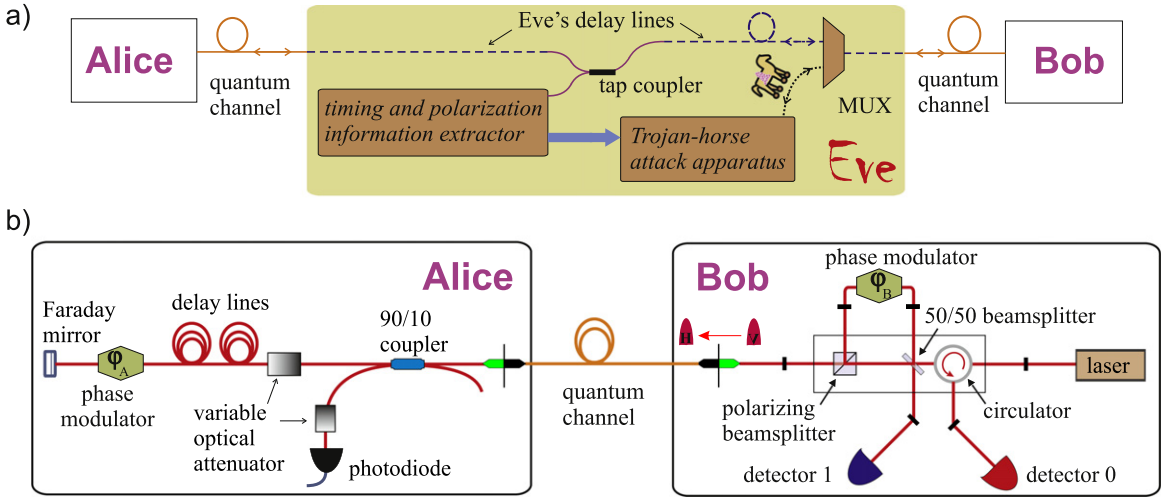

Figure 1. Basic optical schematic of the Trojan-horse attack and plug-and-play QKD system. (a) Using the MUX, Eve multiplexes (in time and wavelength) the Trojan-horse pulses to the quantum signals traveling from Alice to Bob for probing Bob's basis choice. Reflections from Bob travel back to the Trojan-horse attack apparatus after being demultiplexed at the MUX. Eve may also replace parts of the quantum channel (in solid-orange) with her own delay lines (in dashed-blue). (b) A folded Mach–Zehnder interferometer operating in double pass facilitates a passive autocompensation of optical fluctuations (arising in the quantum channel) and forms the essence of plug-and-play schemes. Bob contains both the laser and single-photon detectors connected to his *local* interferometer by means of a polarizing beamsplitter 50/50 beamsplitter, and circulator (henceforth referred to as the PBS-BS-C assembly). Alice employs a Faraday mirror to reflect back the signals sent by Bob. The small black rectangles in Bob denote a pair of FC/PC connectors inside a mating sleeve.

[17]. A frame in our Clavis2 system is configured to be 215 μs long, while the inter-frame separation depends on the total distance between Alice and Bob[4].

2.1. Time of launching the Trojan-horse pulse

Eve launches a Trojan-horse pulse (THP) into Bob at time $t_{E\to B}$ chosen so that the onward pulse and/or one of its back-reflections (from some component or interface inside Bob) travel through Bob's PM while he is applying a voltage on it. As will be explained below, the back-reflected pulse coming out from Bob onto the quantum channel then carries an imprint of whatever random phase shift φ_B had been applied by Bob. The time $t_{E\to B}$ is of course relative to events inside Bob repeating at $f_B = 5$ MHz. To be synchronized to the clock in Bob, Eve may steal a few photons from the bright pulses traveling to Alice using a tap coupler, as shown in figure 1(a). She can extract information such as timing and polarization from the measurement of these photons and use it in the preparation of the THPs.

[4] Lower bound is provided by the delay line in Alice, which for our system results in ~235 μs.

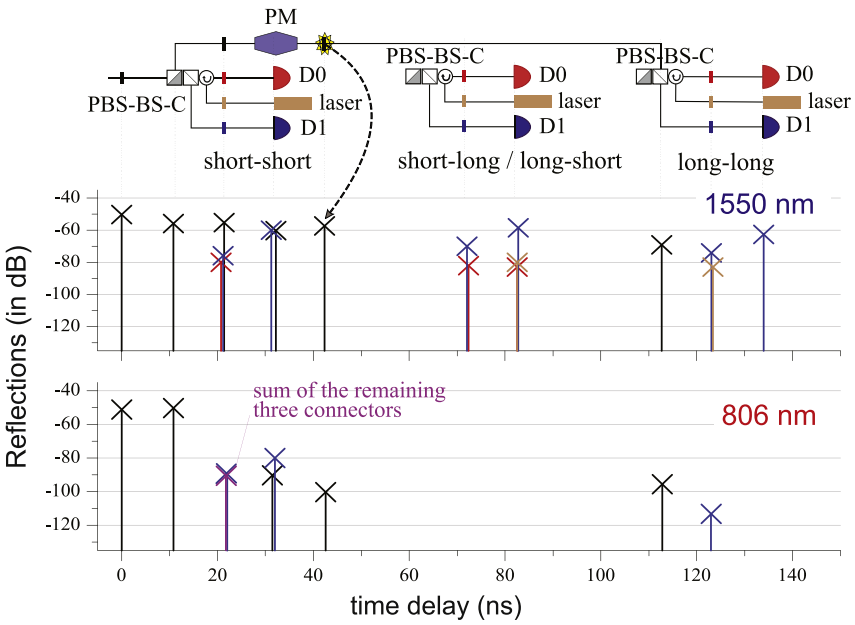

Figure 2. Reflection maps of Clavis2-Bob at 1550 and 806 nm, as seen from Bob's entrance. Reflections from several components close in time are color-coded. Reflections not shown were below the OTDR sensitivity (about −83 dB at 1550 nm and −96 dB at 806 nm). However, some important reflections below the sensitivity limit at 806 nm were estimated by combining several measurements on parts of Bob. The reflection level of the connectors could depend significantly (maximum variation: 3 dB) on the cleanliness of the connectors and mating sleeves. In the scheme, small filled rectangular blocks represent FC/PC connectors with curved polished surfaces; PM: phase modulator, D0 and D1: single-photon avalanche diodes; PBS-BS-C: optical assembly of polarizing BS, 50/50 BS, and circulator. OTDR model: opto-electronics modular picosecond fiber-optic system.

2.2. Time of arrival and amplitude of the back-reflected pulse

As illustrated in figure 1(b), Bob comprises of a miscellany of fiber-optical components. This offers several interfaces from where (measurable) back-reflections could arise. Also, due to the asymmetric interferometer, there may be two different paths traversable in either directions, i.e., for the arrival of the THP into Bob, and departure of a given reflection to the quantum channel. In essence, for a single THP sent into Bob, multiple reflections varying in time and amplitude can be expected. By means of repetitive measurements, a reflection-map for Bob—temporal distribution of the back-reflection levels—can be constructed. This is a task perhaps best suited for an optical time domain reflectometry (OTDR) device [28]. We obtained OTDR traces, or reflection-maps for Bob, for three different wavelengths: 806, 1310 and 1550 nm. Figure 2 illustrates two of them; the traces for 1310 and 1550 nm were found to be quite similar. Due to the polarizing beamsplitter at Bob's entrance (the PBS in the PBS-BS-C assembly), most of the reflection levels depend greatly on the polarization of the probe light. This polarization was set to maximize the reflection from the closest connector of the PM (see star-like shape). As indicated, the back-reflected pulse would exit Bob around 43 ns after the arrival of the THP into Bob; $t_{B\to E} - t_{E\to B} \sim 43$ ns. The corresponding back-reflection level is around −57 dB. By sending a THP, say with a mean photon number $\mu_{E\to B} = 2 \times 10^6$, Eve would get a back-reflection $\mu_{B\to E} \approx 4.0$, i.e., with just four photons on average.

2.3. Measurement of the back-reflected pulse

Per se, any physical property in the back-reflected pulse that provides a clue of Bob's modulation suffices, and governs Eve's measurement technique. If Eve uses a coherent laser operating at wavelength λ_E to prepare the THP, the state of light in the back-reflected pulse can be approximated by a weak coherent state $|\alpha\rangle$. The phase φ_E of this state (relative to a reference) depends on λ_E, e.g., if $\lambda_E = \lambda_{AB} \sim 1550$ nm, and Eve launches the THP so that both the onward and back-reflected pulse make a pass through the PM while it is active, then $\varphi_E \cong 2\varphi_B = 0$ or π. The objective then simplifies to discriminating between two weak coherent states having the same amplitude $|\alpha|$ but opposite phase, which can succeed with a probability $1 - e^{-|\alpha|^2}$ at most (which is the probability that the state $|\pm\alpha\rangle$ is not projected onto the vacuum state). Assuming the aforementioned case with $|\alpha|^2 \equiv \mu_{B\to E} \approx 4.0$, the maximal success probability is 98.2%. The phase reference to probe whether $\varphi_E = 0$ or π may either be a bright local oscillator of a homodyne detector, or an attenuated coherent state (the same level as $\mu_{B\to E}$) and a pair of single-photon detectors.

2.4. Avoiding discovery by Bob (or Alice) and other constraints

Raising $\mu_{E\to B}$ would yield more photons for the measurement, allowing for a better phase discrimination, but how do these bright pulses affect the other components in the QKD system in general? An oddly-behaving component is a signature that could lead to Eve's discovery, so this issue is quite central to the success of Eve's attack.

Bob uses a pair of single-photon avalanche diodes (SPADs) operated in gated mode[5] to detect the legitimate photonic qubits from Alice. Eve's bright pulses, even if timed to arrive outside the detection gate, tend to populate carrier traps [7, 29] in the SPAD. This ensues in an afterpulsing effect: traps exponentially decay by releasing charge carriers that may stimulate avalanches of current, or *afterpulses*, in the onward gates. These afterpulses increase the dark count rate, i.e., result in a higher number of false clicks in the SPADs. Due to this, the quantum bit error rate (QBER) incurred by Alice and Bob at the conclusion of the key exchange will naturally be higher. Eve's objective is to make sure that the QBER does not cross the 'abort threshold' (e.g., around 8% in Clavis2 [8]) as that would fail her eavesdropping attempt. Moreover, as characterized in the so called after-gate attack [7], if the brightness $\mu_{E\to B}$ exceeds a certain threshold, then for a THP arriving a few ns after the gate, the SPAD may register a click with high probability for that particular slot. Since Eve wants to *merely* read the state of the PM via a THP, she must constrain the brightness of this pulse to avoid an undesired click in Bob's SPADs in the attacked slot. This imposes an upper limit on $\mu_{E\to B}$, which is $\sim 2 \times 10^6$ for our system [7].

Since the afterpulsing is strongly dependent on the brightness $\mu_{E\to B}$, Eve would like to attack with the dimmest-possible THPs. The lower limit is mainly decided by the success probability in discerning Bob's modulation as $\mu_{B\to E}$ falls in the few-photon regime. Reducing the frequency of attack $f_{E\text{att}}$ also decreases afterpulses but implies that Eve probes only a fraction of the slots: she can then obtain only a partial knowledge of the raw key. This must therefore be high enough to ensure a positive leakage of information at the end of the protocol,

[5] Gate width for Clavis2 system is ≈2.0 ns, and gate period is $1/f_B = 200$ ns.

i.e., after Alice and Bob have distilled the secret key by estimating Eve's information and destroying it by means of privacy amplification.

2.5. Suitable wavelength for attack

The properties of most optical components, such as attenuation through fibers or back-reflectance of connectors, varies with wavelength. The notable differences between the OTDR traces at 808 and 1550 nm, shown in figure 2, is a testimony to this fact.

Ideally speaking, to characterize a QKD system, one should perform individual OTDR measurements over a large spectral range that could prove feasible for mounting Trojan-horse attacks. However, identifying such a range is not easy. Moreover, it requires an OTDR system with a tunable source as well as a detector with a high sensitivity over the complete range. This may not be possible in practice.

Alternatively, one may spectrally characterize individual optical/optoelectronic components that could play a critical role in preventing or (inadvertently) helping Trojan-horse attacks. To that end, we recently made some spectral measurements on components such as optical isolators with broadband sources to analyze the risks of Trojan attacks at wavelengths other than 1550 nm [13]. We also made some simple measurements on the Bob subsystem to examine the spectral behavior of the PM (in conjunction with its input and output connectors) and the sensitivity of the SPADs. Fortunately for Clavis2, we did not find any reflection peaks at wavelengths far from 1550 nm that could have aided Eve in the attack. Based on figure 2 here and figures 6 and 7 in [13], the optimum attack wavelength seems to be ∼1550 nm.[6]

3. Phase readout experiment

3.1. Eavesdropping setup

Here we describe our implementation of a proof-of-principle Trojan-horse attack. Figure 3 shows the schematic of the apparatus used for reading out the unknown phase by means of homodyne detection. For this, we disconnected Bob from Alice. A pulse & delay generator (Highland Technology P400) was synchronized to Bob and drove Eve's laser at a repetition rate $f_{E\mathrm{att}} = 5$ MHz. An optical isolator was employed to protect Eve's laser from reflections. Using a 1/99 coupler, the THPs were directed into Bob from port 3. The polarization of these THPs was optimized using PC1 so that the power at the FC/PC connector (port 9, inside Bob) after the PM was maximum.

A long fiber patchcord of an appropriate length was spliced and added to the other arm of the coupler at port 4. The relative path difference between the back-reflected pulse (*signal* path) and the local oscillator pulse (*control* path), as observed at the 50/50 BS of the homodyne detector, was adjusted to achieve the maximum interference visibility. The polarization of the *signal* (*control*) pulses at the outcoupler FC1 (FC2) could be controlled by PC2 (PC3). Using P400, the laser delay, i.e., $t_{E\to B}$ was changed so that the input pulse traveled through Bob's PM while the PM was activated. The optical pulse width, and therefore the mean photon number per pulse, could be fine-tuned by changing the driving pulse width in P400.

[6] Although the OTDR traces at 1310 nm and 1550 nm are similar, we believe the afterpulsing induced in Bob's SPADs due to a Trojan-horse attack at 1310 nm may be worse than at 1550 nm.

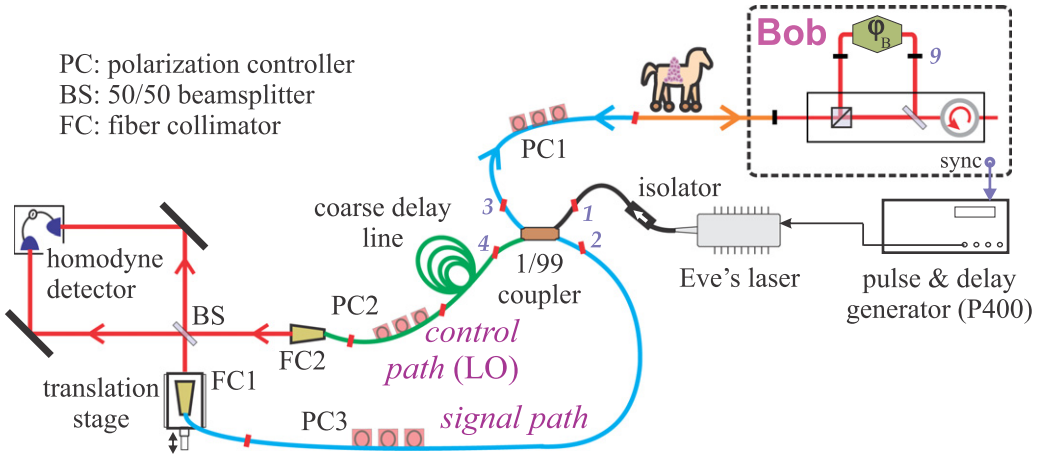

Figure 3. Schematic of a Trojan-horse eavesdropper. Some components in Bob are not shown to avoid cluttering. To synchronize to Bob's modulation cycle, we used an electronic sync signal as shown. In an actual attack, Eve can use the method explained in section 2.1 (also see the explanation of the attack strategy in the appendix).

3.2. Results

We adjusted the power of the laser so that $\mu_{\mathrm{LO}} > 10^8$ and $\mu_{E\to B} \leqslant 1.5 \times 10^6$ (resulting in a mean photon number $\mu_{\mathrm{sig}} \approx 3$ for the homodyne detection) was obtained. We separately confirmed that a slot attacked with such a THP never experienced a click (except due to a dark count) [7]. As mentioned before, Clavis2 operates the quantum key exchange in *frames* that are 215 μs long, containing $N_{\mathrm{f}} = 1075$ modulations or slots repeating every 0.2 μs. We configured the oscilloscope to capture the output voltage of the homodyne detector and the PM voltage (obtained via an electronic tap placed inside Bob) in a single-shot acquisition mode lasting 250 μs.

Figure 4(a) shows the time traces of Bob's randomly-chosen phase modulations and the output of Eve's homodyne detector for five arbitrarily chosen slots. A direct discrimination may not be evident by eye, however, after integrating the homodyne pulses over a suitably chosen time-window every slot, the result illustrated in figure 4(b) is obtained. Figure 4(c) shows these integrated values for an entire frame with the adjacent table showing the number of instances of success/failure in discrimination. In (501 + 518)/(528 + 547) = 94.8% of all the slots, Eve's bits match with those of Bob.

In general, we find that Eve can discriminate >90% Bob's secret bits in all the processed frames. Note that this success rate is conditioned on choosing an appropriate threshold (black horizontal line) which may vary on a frame-to-frame basis due to global phase drifts. However, Eve can always set a reference to $\varphi_B = 0$ by simply crafting the LO pulse train to detect another back-reflected pulse that did a double pass through Bob's PM when it was *in*active; see sections 2.1 and 2.2. In section 5 we shall discuss a few techniques that can increase the discrimination probability in practice. To simplify our simulation, we assume from here on that a THP with $\mu_{E\to B} \sim < 2 \times 10^6$ can always accurately read the state of Bob's PM in each slot.

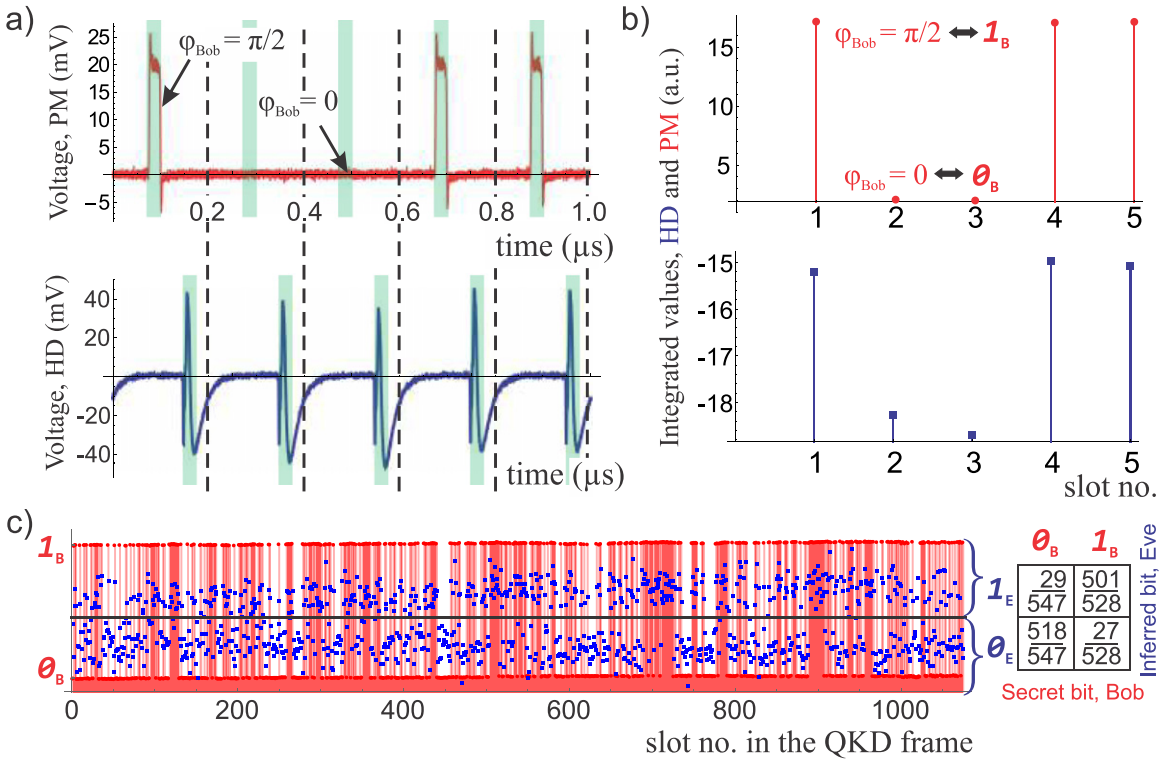

Figure 4. Results of phase readout. (a) Traces of Bob's randomly-chosen phase modulation (in red) and the output of Eve's homodyne detector (in blue) for a sequence of five arbitrarily chosen slots. The measurement was performed at $\mu_{\text{sig}} \approx 3$ and $f_{E\text{att}} = 5$ MHz. The correlation between Bob's phase modulation and Eve's homodyne pulses can be extracted by integration over a time-window (denoted by green shaded rectangle). (b) Thus, in each 200 ns slot, a single value each for the random phase modulation and homodyne pulse is calculated. (c) Series of $N_{\text{f}} = 1075$ such integrated values, scaled and shifted by arbitrary constants merely to aid visual discrimination. Using an appropriate threshold (horizontal black line), Eve's estimation of Bob's secret bit 0_B or 1_B is correct for most of the slots in this frame; see the tabulated values on the right side with bits 0_E and 1_E quantizing Eve's estimation.

4. Eve's attack strategy simulation

To know the entire modulation sequence in Bob, Eve would have to attack the QKD system with $f_{E\text{att}} = 5$ MHz which would result in a tremendous amount of afterpulsing in Bob's SPADs even when $\mu_{E\to B} \sim 2 \times 10^6$ is chosen. A straightforward attack is clearly not possible. In this section, we devise an attack strategy that may still allow Eve to probe Bob's PM frequently enough to obtain a higher percentage of the key than Alice and Bob estimate her to possess during the calculation of the secret key fraction [19]. Neither is the expected detection rate of Bob severely affected, nor does the QBER cross the abort threshold. In other words, *a non-zero portion of the final secret key is leaked to Eve without her being discovered.*

To motivate the basic idea of the strategy, note that it makes sense to probe the modulation in a slot if Bob, with a high probability, eventually obtains a valid detection in that slot. Conversely, if a slot has a very low probability of being registered by Bob, probing that slot is

not only a waste but also the afterpulsing—due to Eve's bright pulses—unnecessarily increases the QBER. By manipulating the photonic frame, i.e., the train of $N_f = 1075$ legitimate weak coherent pulses (WCPs) returning from Alice to Bob, Eve can control the timings of detection events in Bob. For this purpose, she may either (i) use a low-loss channel to transfer the photon (s) in a WCP from Alice to Bob and increase the chance of a click in that given slot, or (ii) block the WCP entirely to decrease it. She multiplexes THPs on (a subset of) the former slots as depicted in figure 1(a) while keeping her laser shut in the latter slots.

Since the mean photon number of the WCPs arriving in Bob is rather low, a major chunk of the slots would actually contain 0 photons, and obviously cannot result in a detection event in Bob. Eve may increase her chance of attacking a slot, that eventually yields a valid detection event, by sending a set of consecutive THPs, here called an *attack burst* with length N_{ab}. However, this burst would also cause a large amount of afterpulsing—noticeable even a few slots after its application. Eve's remedy to this is based on the fact that a successful click causes a *deadtime* in Bob's SPADs. During the attack burst, Eve therefore tries to impose a deadtime in Bob from Alice's photons to *mask* the afterpulsing. To achieve that, she uses the low-loss channel to transfer the N_{ab} slots to Bob to increase the photon detection probability.

Since N_{ab} can obviously not be too large, the deadtime imposition (resulting in $N_{dt} = 50$ gates to be withdrawn) may not always work during the attack burst. Therefore, Eve also transfers another set of N_{ss} slots, called the *substitution sequence*, on the low-loss channel to keep the photon detection probability high after the attack burst. We emphasize that no THPs are added in the substitution sequence.

In this scenario, the detection clicks in Bob's SPADs due to Alice's photons (sent over the low-loss channel in $N_{ab} + N_{ss}$ slots) compete with those from the afterpulses: the former may mask the latter, effectively lowering the error probability. Finally, another optimization for Eve would involve drastically decreasing the detection probability before these $N_{ab} + N_{ss}$ slots—otherwise, a click in a slot before the attack burst slots would result in the burst being encompassed in a deadtime, yielding no benefit to Eve. By extinguishing a certain number of the WCPs (denoted as *extinguished length* N_{el}), she may reduce these chances. Thus, her attack pattern can be thought of as a repetition of the triad $\{N_{el}, N_{ab}, N_{ss}\}$, as illustrated by an example in figure 5(a).

4.1. Evaluating the QKD frame manipulation

In appendix, we describe a specific construction of Eve's strategy using fast optical switches [30] and low-loss channels for manipulating the QKD frame as explained above. Due to this manipulation, Bob receives photons from Alice only during the attack bursts and substitution sequences. This is apparent in figure 5(b); see the thick yellow and green segments. Also, due to the afterpulses emanating from the attack burst slots, the dark noise is not uniform throughout the frame. The overall noise probability in the lth slot is given by $n_j(l) = d_j + a_j(l) - d_j \times a_j(l)$ for $j = 0$ and 1, and is shown in figure 5(c). In this expression, $d_{0/1}$ represents the dark noise probability per gate for D0/D1. The function $a_j(l)$ is computed by summing together the contributions of all previous afterpulses until the lth slot; this is explained in more detail in [7]. Table A1 lists all the parameters for calculating the function $n_j(l)$.

After considering both the photonic input and noise figure, we can evaluate the final detection probabilities $p_j(l) = s_j(l) + n_j(l) - s_j(l) \times n_j(l)$ for the entire frame, as shown in figure 5(d). We explain the derivation of $s_j(l)$ and modeling of the click events in D0 and D1

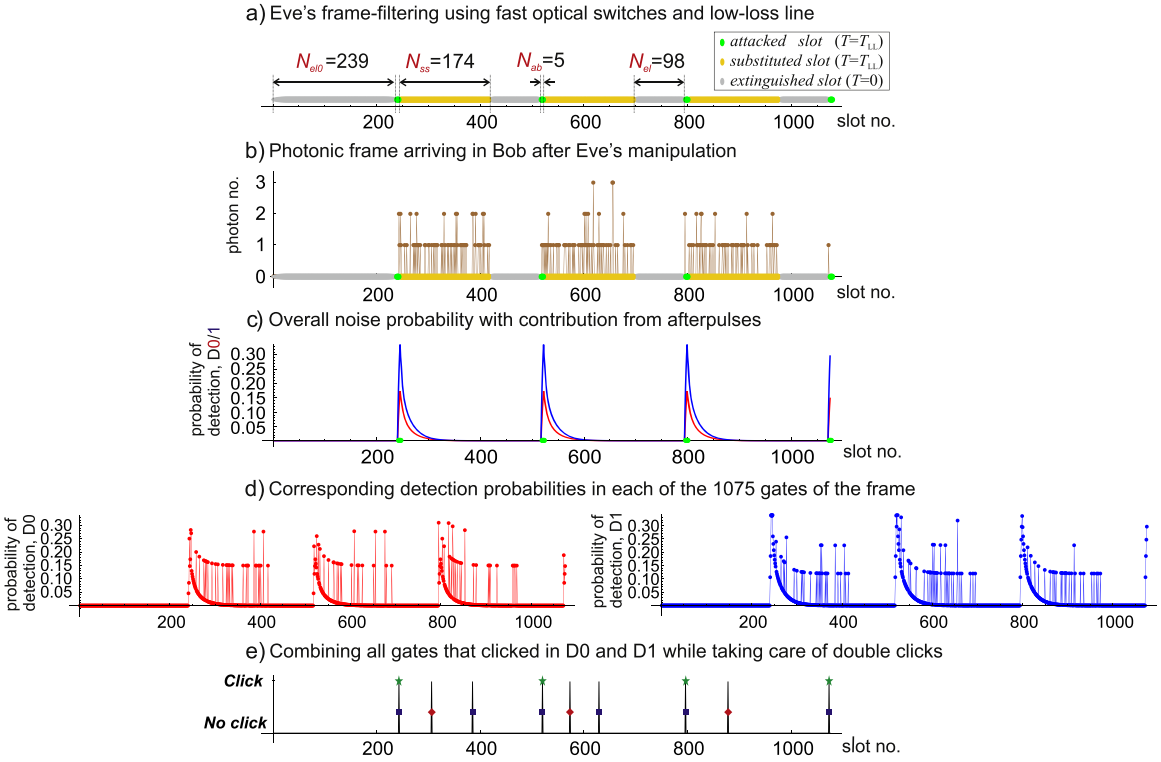

Figure 5. Simulating the effect of Eve's attack on the QKD protocol operation. (a) Eve manipulates a frame sent by Alice to Bob using the strategy described in the main text (more details in the appendix). (b) Bob receives a 'filtered' frame, as the *effective* channel transmission is $T = T_{\rm LL}$ for all $N_{\rm ab}$ (attack burst) and $N_{\rm ss}$ (substitution sequence) slots, and $T = 0$ for $N_{\rm el}$ (extinguished length) slots. We assumed $T_{\rm LL} = 0.9$ in the present case. (c) Characteristic exponential decay of probabilities due to afterpulsing in both D0 and D1 (red and blue) may be visualized after the attack-bursts. (d) Final detection-probability patterns for D0 and D1. (e) Subsequent click pattern just like in figure A1(e); out of the 9 slots (3 in D0 and 6 in D1, indicated by rotated-red and straight-blue squares, respectively) where clicks occurred, Eve knows the basis choice of Bob in 4 of them (indicated by green star).

based on Bernoulli trials in appendix. Figure 5(e) illustrates the clicked gates found after taking double clicks and deadtime imposition into account. Note that while Eve attacked only 20 out of 1075 slots, she knows Bob's basis choice in 4 out of 9 slots that are going to be used in the formation of the raw key.

The QBER incurred by Alice and Bob is strongly dependent on the combination $\{N_{\rm ab}, N_{\rm ss}, N_{\rm el}\}$ used by Eve during the operation of the QKD protocol. The quantum channel transmission T and low-loss line transmission $T_{\rm LL}$ directly influence the photon number statistics $\mu_{\rm SARG04}$ in Alice and the observed detection rate γ_B in Bob, and also indirectly affect both the QBER and Eve's actual knowledge $I_E^{\rm act}$ of the key shared by Alice and Bob after error correction (EC). For instance, long and frequent attacks (larger $N_{\rm ab}$ and smaller $N_{\rm ss}$, in a relative sense) yield high $I_E^{\rm act}$ but also high QBER. Similarly, a large $N_{\rm el}$ preceding an attack burst may effectively increase $I_E^{\rm act}$ as the attacked slots have lesser chances of being inside a deadtime period, but this may also decrease γ_B. And a high $T_{\rm LL}$ naturally implies higher γ_B, and perhaps lower QBER because the dark noise is effectively decreased, however $T_{\rm LL}$ cannot exceed 1.

4.2. Classical processing and optimizing the simulation

Let us first briefly recapitulate some essential information from the previous pages. In section 3, we experimentally demonstrated the readout of Bob's PM with a high accuracy. However, we also found that frequent THPs would result in a huge afterpulsing in Bob's SPADs which would reveal Eve's presence easily. In this section, we devised an intuitive strategy in which Eve manipulates the frame-based communication of Clavis2 and attacks (with THPs) only a small but carefully-chosen subset of the slots in a frame. If Eve simultaneously ensures that

(i) the QBER q does not cross the abort threshold ($q < q_{\text{abort}}$),

(ii) the portion of the error-corrected key Eve actually knows is more than whatever Alice and Bob estimate based on the security proof ($I_E^{\text{act}} > I_E^{\text{est}}$), and

(iii) the deviation of the observed detection rate γ_B^{obs} from the expected value in Bob γ_B^{exp} , given by $\delta_B = \left| 1 - \frac{\gamma_B^{\text{obs}}}{\gamma_B^{\text{exp}}} \right|$, is within tolerable limit ($\delta_B \leqslant \delta_B^{\max}$),

then her strategy succeeds. For satisfying these requirements, one needs to find an optimal attack combination. We simulated different combinations $\{r, N_{\text{ab}}, N_{\text{ss}}, N_{\text{el}}\}$; with the new variable $r \leqslant 1$ denoting the fraction of frames subjected to the Trojan-horse attack. To elaborate, if $r = 0.8$, Eve randomly chose 80 out of 100 frames to attack with the pattern imposed by a specific triad $\{N_{\text{ab}}, N_{\text{el}}, N_{\text{ss}}\}$ in the manner shown in figure 5, while the remaining 20 passed to Bob normally (in the manner shown in figure A1).

Due to probabilistic elements in the simulation, each run was performed for $n_{\text{sim}} = 10\,000$ frames to minimize stochastic fluctuations. In each run, slots that yielded clicks were collated and the average number of clicks per frame γ_B^{obs}= (total clicks)/n_{sim} was calculated. A basis reconciliation procedure, as per the specifications of SARG04 [18, 19], was then performed on the collated slots. This provided us with the incurred QBER q and the fraction of *valid* slots[7] in which Eve knows the secret bit. From the former, we can calculate the leak due EC $leak_{\text{EC}}$ then use it with the latter to bind Eve's knowledge I_E^{act} of the error-corrected key. In particular, we assumed EC to work in the Shannon limit, i.e., $leak_{\text{EC}} = h(q)$, with $h(x) = -x \log_2(x) - (1 - x) \log_2(1 - x)$ being the binary entropy.

To calculate the amount of privacy amplification that Alice and Bob do in SARG04 protocol, we evaluated the expression $I(A{:}\,E)$ derived in [19] (equation (88) therein); this provides I_E^{est} essentially. The derivation considers eavesdropping strategies applicable against SARG04 when Alice employs an attenuated laser instead of a single-photon source. The final expression is obtained while optimizing and lower-bounding the secret key fraction attained by Alice and Bob.

One element considered in the calculation of $I_E^{\text{act}} = I(A{:}\,E)$ is *preprocessing*: a classical operation performed by Alice at the commencement of QKD that reduces both Bob's and Eve's information, but in a more inimical manner for the latter than the former [19, 31]. Although [19] concludes that preprocessing in SARG04 helps Alice and Bob only in a very specific regime, it does not explicitly state that preprocessing should be avoided in other regimes. Since security proofs generally consider attacks that maximize $I(A{:}\,E)$ instead of $I(B{:}\,E)$, the use of preprocessing by Alice may expose a vulnerability exploitable via Trojan-horse attacks on Bob.

[7] That is, the slots kept by both Alice and Bob after the basis reconciliation.

Although preprocessing is not implemented in Clavis2, we consider a case here to highlight the vulnerability.

Indicating the degree of preprocessing with y, and using all the relevant source, channel, and detector parameters introduced thus far, we calculate $I_E^{\text{est}} = 0.4844$ for $y = 0$. This implies that Alice and Bob compress almost half of their error-corrected key during privacy amplification. If however, Alice were to use the maximum preprocessing ($y = 0.5$), then $I_E^{\text{est}} = 0.1106$. Note that the value of I_E^{est} is independent of the incurred QBER. This is due to the fact that the attacks found optimal in the security proof [19] are 'zero-error' attacks [3]. However, I_E^{est} depends on the channel transmission, as also shown in [19]. The values here are calculated at a fixed transmission ($T = 0.25$).

5. Results and discussion

We found several optimal combinations $\{r, N_{\text{ab}}, N_{\text{el}}, N_{\text{ss}}\}$ that satisfy two of the three conditions listed in the previous section ($q_{\text{abort}} \approx 0.08$ and $\delta_B^{\max} = 0.15$ for Clavis2). These are shown in figure 6(a). However, it is clear that $I_E^{\text{act}} << I_E^{\text{est}}$, i.e., Eve's knowledge never surpasses the estimate of Alice and Bob. The reason for the failure is that the detectors, especially D1, in Clavis2 are quite noisy: even without an attack, i.e., with $r = 0$, the QBER $q = 2.52\%$. Crafting an attack with high r and optimal $\{N_{\text{ab}}, N_{\text{ss}}, N_{\text{el}}\}$ may give Eve sufficiently high I_E^{act} but the incurred QBER $q >> q_{\text{abort}}$.

If we assume Bob's detectors to have the same characteristics as that of D0 (in Clavis2), and that Alice has preprocessing accidentally enabled, then Eve could breach the security for $q_{\text{abort}} \approx 0.11$ as shown in figure 6(b). This is possible because the mutual information between Eve and Bob scales by the same factor (given by $1 - y$) as that between Alice and Bob: in particular, at $y = 0.4$, Eve can surpass $I_E^{\text{est}} = 0.1336$.

In order to gauge the full power of this attack strategy and the dangers posed by Trojan-horse attacks in general, we optimized the simulation for a Clavis2-like QKD system assumed to be fitted with a pair of SPADs having *high efficiency* and *low noise*. To be more precise, we assumed a pair of gated SPADs with detection efficiencies $\eta_0 = \eta_1 = 0.25$, thermal dark count probabilities $d_0 = d_1 = 10^{-5}$ per gate, and a cumulative probability of obtaining random click after deadtime period due to afterpulses to be <10% (refer to table A1 for comparison). Note that detectors with similar or even better characteristics have already been reported [32–35], thanks to the recent advances in single-photon detection technology. Alternatively, mechanisms to photoionize the trapped charges through sub-band energy illumination in order to reduce afterpulsing have also been investigated [36]. Therefore, it is quite reasonable to expect such characteristics in the next-generation gated SPADs in Clavis2 or recently-manufactured QKD devices. In such QKD systems, not only can Eve attack more often, but also expect detections from photons to exceed those from afterpulses.

Figure 6(c) shows some optimized attacks ($I_E^{\text{est}} = 0.5037$ for the new detector parameters and no preprocessing) that satisfy all the three conditions. In particular, the positive leakage $I_E^{\text{act}} - I_E^{\text{est}}$, which is likely to be higher when preprocessing is also used, implies that the security of the QKD system would be breached.

At lower channel transmission values ($T < 0.25$), attack regimes with a positive leakage of the final secret key may be found by means of more exhaustive optimization of the simulation.

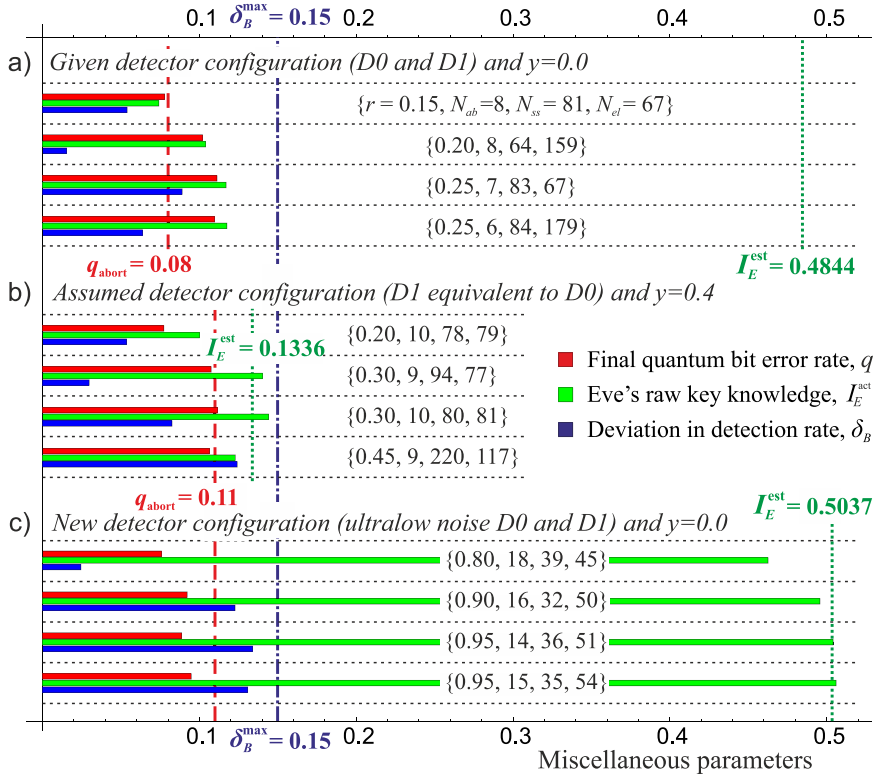

Figure 6. Performance of the simulated attack strategy in three different scenarios. The QKD system aborts the protocol when the QBER q crosses a threshold q_{abort} (dashed red line) or the absolute deviation in the detection rate δ_B surpasses a boundary $\delta_B^{\max}$ (dash-dotted blue line). To break the security under these constraints, Eve's actual knowledge of the key I_E^{act} must exceed the estimate made by Alice and Bob I_E^{est} (dotted green line). (a) Assuming D0 and D1 with characteristics as that of the Clavis2 detectors (see table A1) and that Alice does not apply any preprocessing ($y = 0$), it seems difficult to satisfy the three conditions: $q < q_{\text{abort}}$, $\delta_B \leqslant \delta_B^{\max}$, and $I_E^{\text{act}} > I_E^{\text{est}}$ simultaneously. (b) Assuming both detectors are behaving like D0, some preprocessing ($y = 0.4$), and $q_{\text{abort}} \approx 0.11$, Eve can breach the security. (c) A QKD system implemented with SPADs having high efficiency and low noise is vulnerable to Trojan-horse attack even without the preprocessing loophole. The optimal attack combinations $\{r, N_{\text{ab}}, N_{\text{ss}}, N_{\text{el}}\}$ that produced these results are also listed (see text for details). All parameters and results were computed at $T = 0.25$ and $T_{\text{LL}} = 0.9$.

At $T > 0.25$, Eve's attack should have better chances of succeeding because Alice's quantum states have more photons on average, which raises the photonic detection probability (effectively suppressing the afterpulsing probability) in Bob. However, the calculation of I_E^{est} in the security proof [19] is valid only for channel lengths above 24 km, translating roughly into $T < 0.33$. More photons from Alice also raise the chances of better photon-number-splitting attacks [9, 21] which would require increasing I_E^{est} in privacy amplification, thereby requiring Eve to work harder.

Nonetheless, it is clear that our attack on a QKD system equipped with less noisy SPADs would succeed at least for a range of channel transmissions. A finite amount of preprocessing—supposed to provide *more* security to Alice and Bob—would actually relax the constraints on

Eve. Finally, this strategy could be combined with other hacking strategies, such as the after-gate attack [7], to enhance Eve's performance.

5.1. Possible improvements and extensions

An optimization over the complete space of all parameters that define the attack strategy is out of the scope of this work, but a powerful adversary can easily do so and is likely to find a new set of parameters with better attack performance. A possible extension of the strategy is to manipulate the frames from Bob to Alice as well: more precisely, to replace the legitimate bright pulses in the slots chosen for the *attack burst* with even brighter ones. This would increase the chances that these slots eventually yield valid detections in Bob. Unfortunately, an increased optical power, even if only for a few pulses in the frame, portends a risk for Eve because the monitoring detectors in Alice may raise an alarm. However, if the monitoring system in Alice either does not function properly, or can be fooled [14], then this method holds a lot of promise.

Yet another attack optimization is non-demolition measurement [37] of the photon numbers of the WCPs exiting Alice. Using it, Eve can simply withhold her attack in the slots that contain 0 photons. This would reduce the dark counts (from afterpulsing), yet effectively increase her knowledge of the key. Finally, with regards to the attack setup shown in figure 3, Eve could:

- gather more information (per phase modulation) by suitably tweaking her LO to homodyne *multiple* back-reflections and improve the quality of the phase readout,
- periodically track the phase drift in her setup and adjust the relative phase between the signal and LO, e.g., by using an extra PM in the LO arm, to always read out at an optimal phase difference, and/or
- enhance the success rate of discrimination by using better quantum measurement strategies [38] and post-processing techniques, e.g., taking the difference of consecutive pulses and then integrating over the properly-chosen time window.

These methods would increase Eve's discrimination probability to ∼100% (see figure 4) while relaxing the brightness requirement, i.e., $\mu_{E\to B}$ may be lowered, thus bringing down the afterpulsing probability. Another way to achieve the same goal would be to employ longer wavelengths to attack (as the afterpulsing response of the SPADs is conjectured to be lower) and/or to depopulate the traps by means of photoionization. Eve could try to use ∼1700 nm for her THPs to reduce afterpulsing. A CW illumination at a longer wavelength ∼1950 nm may depopulate the traps (created due to the THPs at some other wavelength) by means of photoionization [36].

The attack setup shown in figure 3 can be used virtually against any kind of QKD system, including CVQKD devices [23, 24]; it only needs a careful delay and polarization control and interferometric stability. It can even be made portable by integrating a variable optical delay line, a low-loss 90° optical hybrid, etc. [27]. Finally, the strategy detailed above can also be attuned to attack entanglement-based QKD systems that may not have proper safeguards against Trojan-horse attacks. More significantly, it may be used even to break the BB84 protocol in such cases.

5.2. Countermeasures

Experimentally speaking, isolators and wavelength filters have been the most suitable countermeasures against Trojan-horse type attacks for one-way QKD systems [12]. While the former cannot be used in a two-way QKD system like Clavis2, the latter can certainly be useful. In a related context, one must also scrutinize (high and unwarranted) back-reflections from the interfaces inside the QKD system that could pose risks as explained in section 2. With such analysis, it might be possible to incorporate Trojan-horse attacks into theoretical security proofs and neutralize them by correct levels of privacy amplification. Moreover, security proofs should also carefully examine and quell the undesired effects of preprocessing. Some technical countermeasures specifically for the Clavis2 system could be:

- installing a watchdog detector with a switch at the entrance of Bob that randomly routes a small fraction of incoming signals to this detector,
- opening the door for Eve for a smaller time duration, i.e., reducing the width of phase modulation voltage pulse, and
- monitoring Bob's SPADs in real time [39].

Except the watchdog detector countermeasure, all others require modifications only in the electronic control system and hence are recommended.

Note that Bob's vulnerability to the Trojan-horse attack only arises because the SARG04 protocol is used. For BB84 (including its decoy-state version), interrogating Bob's modulator gives Eve no advantage [12], except when this is used to counterattack the four-state patch to the detector efficiency mismatch attacks [20, 40]. However both BB84 and SARG04 are vulnerable to interrogating Alice's modulator.

6. Conclusion

In conclusion, we have demonstrated the operation of a setup to launch a Trojan-horse attack on a commercial QKD system from ID Quantique. Our objective is to read the state of the PM in Bob to break the SARG04 protocol. We have shown that this phase readout can be performed in real-time with a high success rate, and analyzed various constraints and problems in mounting a full attack on the system. These arise mainly due to the afterpulsing noise induced in the single-photon detectors of Bob by the bright THPs from Eve. We have devised and numerically modeled an attack strategy to keep the overall QBER (which increases due to the afterpulsing noise) below the abort threshold, while allowing Eve to gain maximum-possible knowledge of the raw key. Although on our Clavis2 system, this does not exceed the theoretical security estimate that Alice and Bob make about Eve's knowledge, we have shown that similar or future QKD systems with less-noisy detectors may facilitate Eve's attack to become traceless. We have also proposed some mechanisms to improve the performance of the attack. With some simple modifications, our attack setup and strategy could be applied against many other quantum cryptographic implementations, including entanglement-based, continuous-variable, and measurement-device-independent QKD systems. Finally, we have proposed both general and specific countermeasures that can be easily adopted in most QKD systems.

Acknowledgments

We would like to thank M Legré from ID Quantique, D Sych, C Wittmann, S Pirandola, and L Lydersen for useful discussions. We also gratefully acknowledge L Meier and A Käppel for their assistance in design of electronics. This work was supported by the Research Council of Norway (grant no. 180439/V30), Industry Canada, DAADppp mobility program financed by NFR (project no. 199854) and DAAD (project no. 50727598). E A acknowledges support from CryptoWorks21. V M acknowledges support from University Graduate Center in Kjeller.

Appendix

A.1. Operation of plug-and-play QKD

Here we simulate the operation of the QKD system. A Clavis2 frame consists of $N_{\mathrm{f}} = 1075$ slots spaced 0.2 μs apart. This implies N_{f} optical signals are sent by Bob to Alice in the forward path of the plug-and-play scheme, N_{f} detection gates are opened by Bob to measure the N_{f} WCPs coming back from Alice. Note that in practice, Bob has an asymmetric interferometer as shown in figure 1(b) so an optical signal actually consists of two (unequally bright) pulses; as it does not affect our analysis, we use 'signal' and 'pulse' interchangeably to keep the explanation simple. Alice attenuates these optical signals properly so that the mean photon number of the WCPs (in the quantum channel) is as dictated by the protocol; for SARG04 the optimal value is $\mu_{\mathrm{SARG04}} = 2\sqrt{T}$, where T is the channel transmission [19].

By means of a Monte Carlo simulation based on experimental parameters, we modeled the frame-based QKD operation from here on. We created an array of random positive integers that are Poisson-distributed to mimic (the photon numbers of) a Clavis2 frame exiting Alice. Each pulse in the frame was stochastically subjected to all the relevant transmission or detection events; to be precise, they were modeled by a sequence of Bernoulli trials. For example, if the transmission of the quantum channel is denoted by T, then each of the n photons in a pulse at Alice's exit undergoes a Bernoulli trial yielding success/1 (failure/0) with a probability of T $[1 - T]$. The total number of photons in a pulse reaching Bob can then be evaluated as the sum of the outcomes of all n trials. Similarly, for a pulse containing m photons impinging on an SPAD with single-photon detection efficiency η, a detection click (success) is obtained if at least one of the m Bernoulli trials yielded a 1.

Table A1 lists the various parameters of the detectors in Clavis2 used for the numerical simulation. Figure A1 charts the different events in Bob: right from the arrival of a photonic frame to the registration of clicks, taking the withdrawal of $N_{\mathrm{dt}} = 50$ gates (due to deadtime) into account. The transmission of the quantum channel connecting Alice and Bob is assumed to be $T = 0.25$ (with channel attenuation $\alpha = 0.2$ dB km^{-1}, this would imply $\sim$30 km long channel). The transmission inside Bob is $T_B = 0.45$. The total detection probabilities in figure A1(c) are calculated using $p_j(l) = s_j(l) + d_j - s_j(l) \times d_j$ for each slot $l \in \left[1, N_{\mathrm{f}}\right]$ and for $j = 0$ and 1. In this expression, $d_{0/1}$ represents the dark count probability per gate for D0/D1. The photonic detection probability is $s_j(l) = 1 - \left(1 - \eta_j\right)^{m(l)}$ for $j = 0$ and 1; here $m(l)$ is the number of photons impinging on a specific detector in the lth slot (shown in figure A1(b)), and η_0 and η_1 are the single-photon detection efficiencies of D0 and D1, respectively.

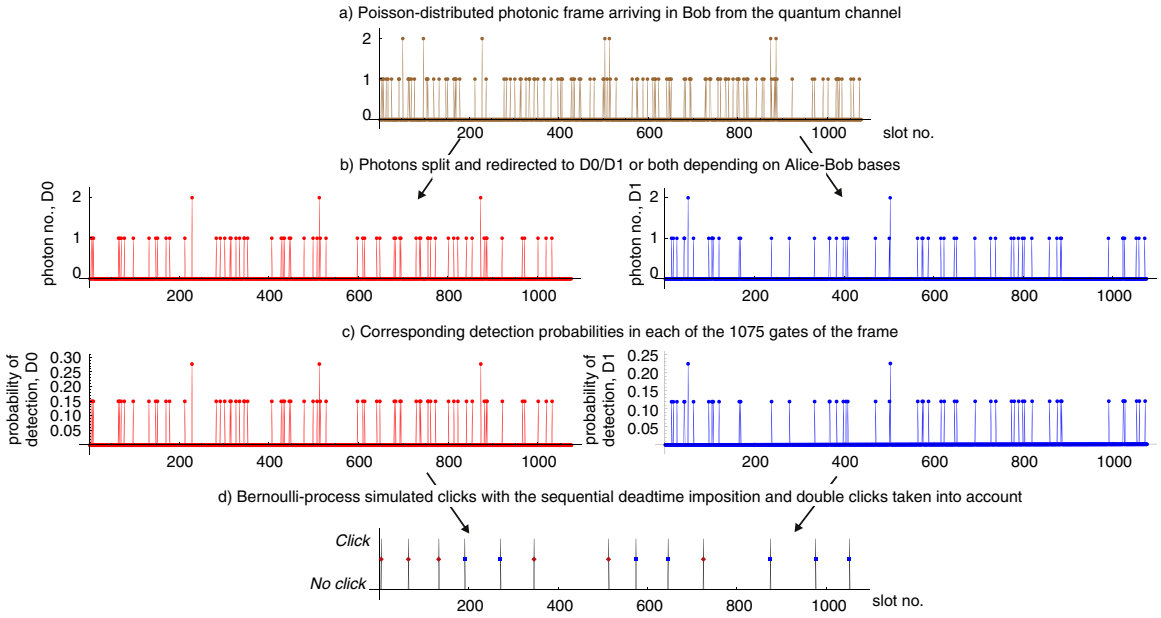

Figure A1. Simulation of the physical-layer operation of SARG04 in Clavis2 at channel transmission $T = 0.25$. (a) Photon number statistics of the WCP train (mean photon number $\mu_{\mathrm{SARG04}} = 1$) that reaches Bob after traversing the quantum channel. In each of the 1075 slots, Alice randomly prepared one of four states $Z0$, $Z1$, $X0$, $X1$. (b) Bob randomly chose Z or X basis in each slot too; if his basis coincides with the preparation-basis of Alice, all photons in that slot are directed to one of D0 or D1 (depending on Alice's secret bit). For dissimilar basis choice, photons are randomly split across D0 and D1. (c) Resultant detection probabilities for D0 and D1 in each slot/gate; calculation details are given in the main text. (d) Subsequent detection-click pattern (vertical black bars with rotated-red or straight-blue squares).

Table A1. Various detection-related parameters in Clavis2. The numerical parameters for the exponential decay due to afterpulses were estimated in [7]. The cumulative probability to get a random click after $N_{\mathrm{dt}} = 50$ gates from afterpulses alone surpasses 80%. The subscript $j = 0/1$ in a variable affiliates it to D0/D1.

	D0	D1
Single-photon detection efficiency, η_j	0.12	0.10
Dark noise probability, d_j	1.16×10^{-4}	3.63×10^{-4}
Afterpulse probability amplitude, A_{1j}	3.572×10^{-2}	10.68×10^{-2}
Afterpulse decay constant, τ_{1j} (μs)	1.159	0.705
Afterpulse probability amplitude, A_{2j}	2.283×10^{-2}	5.054×10^{-2}
Afterpulsing decay constant, τ_{2j} (μs)	4.277	3.866

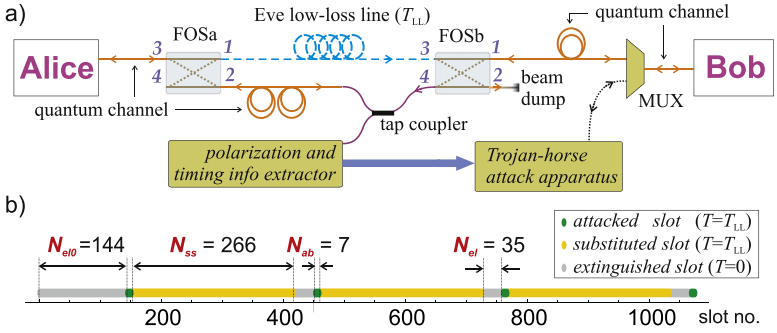

Figure A2. Technical implementation details of the frame manipulation strategy. (a) Eve plants two bi-directional 2 × 2 fast optical switches FOSa and FOSb near Alice and Bob, respectively. The solid orange line represents the quantum channel (normal transmission T) containing an optical tap along with the two switches. The dashed cyan line is Eve's highly-transmissive channel which may be implemented by a low-loss delay line. The operational details of the switches during the quantum key exchange are described in the text. (b) In a frame sent by Alice to Bob, Eve diverts all the slots marked in green and yellow (four sets of N_{ab} and three sets of N_{ss}, respectively) onto a highly-transmissive channel. The slots marked in gray (three sets of N_{el} and one N_{el0}) are blocked. FOS: fast optical switch, MUX: multiplexer, ab: attack burst, ss: substitution sequence, el: extinguished length.

A.2. Eve's strategy

Figure A2 (a) shows a possible full implementation of the Trojan-horse attack described in section 4, by using off-the-shelf optical switches [30] and a low-loss line. The switches are connected by two lines: the quantum channel containing an optical tap additionally, and a highly-transmissive channel (with transmission T_{LL}). If a slot $l \in [1, N_f]$ diverted by Eve on the highly-transmissive channel had n photons at Alice's exit, then it has a high chance of having n photons at Bob's entrance too. The low-loss line with the characteristics we model ($T_{LL} = 0.9$ instead of 0.25 for the normal line) currently does not exist. However, its implementation can in principle be possible in the future, by using an improved optical fiber or high-efficiency quantum teleportation.

When Bob sends a frame to Alice, the switches are in crossed positions (FOSb: 1 → 4 and FOSa: 2 → 3) so that the frame essentially traverses the quantum channel undisturbed. The tap is used for obtaining polarization information and synchronization, required later in preparation of the THPs. Since the pulses in the forward path are relatively bright, a few photons stolen would not be noticed by Alice.

For the return path, i.e., from Alice to Bob, Eve manipulates the slots as determined by the attack pattern of figure A2(b). This pattern is essentially a repetition of the triad $\{N_{ab}, N_{el}, N_{ss}\}$ imposed in the reverse direction (i.e., going from N_f to 1) on an entire QKD frame. The number of unbroken triads that can fit inside a frame is $k = \left\lfloor N_f / \left(N_{ab} + N_{el} + N_{ss}\right) \right\rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor operation. This leaves exactly $N_u = N_f - k\left(N_{ab} + N_{el} + N_{ss}\right)$ *unaccounted* slots in the beginning of the frame; if $N_u > N_{ab}$, then we add yet another attack burst N_{ab} and extinguish the remaining $N_{el0} = N_u - N_{ab}$ slots, as also shown in figure A2(a) with $k = 4$ and $N_u = 151$. Otherwise, we simply extinguish $N_{el0} = N_u$ slots.

Using this pattern, Eve physically manipulates the frame in the following way: slots up to N_{el0} are extinguished by being directed onto a beam dump (FOSa: 3 → 2 and FOSb 4 → 2).

The next $N_{ab} + N_{ss}$ slots pass through the low-loss line (both FOSa and FOSb in positions $3 \rightarrow 1$) to Bob. Using the Trojan-horse attack apparatus (see figure 3), Eve reads Bob's PM settings for the *attack burst*, i.e., the first N_{ab} of these slots. The remaining N_{ss} slots, or the *substitution sequence*, simply travel to Bob via the low-loss line. The switches then flip again for an *extinguished length* of N_{el} slots. This sequence is repeated until the end of the frame is reached with the last N_{ab} gates always attacked. Attacking the last few slots causes less afterpulsing, because the detector gates are not applied after the frame end.

References

[1] Bennett C H and Brassard G 1984 Quantum cryptography: public key distribution and coin tossing *Proc. IEEE Int. Conf. on Computers, Systems and Signal Processing (Bangalore, India)* p 175
[2] Gisin N, Ribordy G, Tittel W and Zbinden H 2002 *Rev. Mod. Phys.* **74** 145
[3] Scarani V, Bechmann-Pasquinucci H, Cerf N J, Dušek M, Lütkenhaus N and Peev M 2009 *Rev. Mod. Phys.* **81** 1301
[4] Nauerth S *et al* 2009 *New J. Phys.* **6** 065001
[5] Lydersen L, Wiechers C, Wittmann C, Elser D, Skaar J and Makarov V 2010 *Nat. Photonics* **4** 686
[6] Li H W *et al* 2011 *Phys. Rev.* A **84** 062308
[7] Wiechers *et al* 2011 *New J. Phys.* **13** 013043
[8] Jain N *et al* 2011 *Phys. Rev. Lett.* **107** 110501
[9] Jiang M S *et al* 2012 *Phys. Rev.* A **86** 032310
[10] Gisin N, Fasel S, Kraus B, Zbinden H and Ribordy G 2006 *Phys. Rev.* A **73** 022320
[11] Bethune D S and Risk W P 2000 *IEEE J. Quantum Electron.* **36** 340
[12] Vakhitov A, Makarov V and Hjelme D R 2001 *J. Mod. Opt.* **48** 2023
[13] Jain N, Stiller B, Khan I, Makarov V, Marquardt C and Leuchs G 2014 *IEEE J. Sel. Top. Quantum Electron.* at press arXiv:1408.0492
[14] Sajeed S, Radchenko I, Kaiser S, Bourgoin J-P, Monat L, Legré M and Makarov V 2014 *Proc. QCrypt (Paris, France)*
[15] Walenta N *et al* 2014 *New J. Phys.* **16** 013047
[16] ETSI GS QKD 005 V1.1.1: Quantum key distribution (QKD); Security proofs (ETSI, 2010)
[17] Stucki D, Gisin N, Guinnard O, Ribordy G and Zbinden H 2002 *New J. Phys.* **4** 41
[18] Scarani V, Acín A, Ribordy G and Gisin N 2004 *Phys. Rev. Lett.* **92** 057901
[19] Branciard C, Gisin N, Kraus B and Scarani V 2005 *Phys. Rev.* A **72** 032301
[20] Makarov V, Anisimov A and Skaar J 2006 *Phys. Rev.* A **74** 022313
[21] Brassard G, Lütkenhaus N, Mor T and Sanders B C 2000 *Phys. Rev. Lett.* **85** 1330
[22] Datasheet of Clavis2, available at ID Quantique website (www.idquantique.com)
[23] Jouguet P *et al* 2013 *Nat. Photonics* **7** 378
[24] Khan I *et al* 2013 *Phys. Rev.* A **88** 010302
[25] Liu Y *et al* 2013 *Phys. Rev. Lett.* **111** 130502
[26] Silva T F *et al* 2013 *Phys. Rev.* A **88** 052303
[27] Khan I *et al* in preparation
[28] Beller J 1998 *OTDRs and Backscatter Measurements in Fiber Optic Test and Measurement* ed D Derickson (Englewood Cliffs, NJ: Prentice-Hall)
[29] Haitz R H 1965 *J. Appl. Phys.* **36** 3123
Cova S, Lacaita A and Ripamonti G 1991 *IEEE Electron Device Lett.* **12** 685
[30] Nanona ultra-fast optical switch (www.bostonati.com)
NanoSpeed (www.agiltron.com)
[31] Kraus B, Gisin N and Renner R 2005 *Phys. Rev. Lett.* **95** 080501

[32] Patel K A *et al* 2012 *Electron. Lett.* **48** 111
[33] Walenta N *et al* 2012 *J. Appl. Phys.* **112** 063106
[34] Restelli A, Bienfang J C and Migdall A L 2013 *Appl. Phys. Lett.* **102** 141104
[35] Korzh B *et al* 2014 *Appl. Phys. Lett.* **104** 081108
[36] Krainak M A 2005 *Proc. Lasers Electro-Opt. (CLEO)* **1** 588
[37] Xiao Y F *et al* 2008 *Opt. Express* **16** 21462
Braginsky V B and Khalili F Y 1996 *Rev. Mod. Phys.* **68** 1
[38] Wittmann C *et al* 2008 *Phys. Rev. Lett.* **101** 210501
[39] Silva T F, Xavier G B, Temporao G P and von der Weid J P 2012 *Opt. Express* **20** 18911
[40] Qi B, Fung C-H F, Lo H-K and Ma X 2007 *Quantum Inf. Comput.* **7** 73 (www.rintonpress.com/xgic7/gic-7-12/073-082.pdf)